Procedia Computer Science

Volume 80, 2016, Pages 2322–2326

ICCS 2016. The International Conference on Computational Science

An Execution Framework for Grid-Clustering Methods

Erich Schikuta[1] and Florian Fritz[1]

University of Vienna, Faculty of Computer Science, RG WST, Vienna, Austria
erich.schikuta@univie.ac.at, a0800640@unet.univie.ac.at

Abstract

Cluster analysis methods have proven extremely valuable for explorative data analysis and are also fundamental for data mining methods. Goal of cluster analysis is to find correlations of the value space and to separate the data values into a priori unknown set of subgroups based on a similarity metrics. In case of high dimensional data, i.e. data with a large number of describing attributes, clustering can result into a very time consuming task, which often limits the number of observations to be clustered in practice. To overcome this problem, Grid clustering methods have been developed, which do not calculate similarity values between the data value each, but organize the value space surrounding the data values, e.g. by specific data structure indices. In this paper we present a framework which allows to evaluate different data structures for the generation of a multi-dimensional grid structure grouping the data values into blocks. The values are then clustered by a topological neighbor search algorithm on the basis of the block structure. As first data structure to be evaluated we present the BANG file structure and show its feasibility as clustering index. The developed framework is planned to be contributed as package to the WEKA software.

Keywords: Cluster Algorithms, Dendrogram, Grid-based Methods

1 Introduction

Cluster analysis methods [3] are key for explorative data analysis in statistics and computer science [9], but are also fundamental for data mining methods [1]. Goal of cluster analysis is to find correlations of the value space and to separate the data values into a priori unknown set of subgroups based on a similarity metrics.

In the past a large number of different clustering approaches have been presented, which are classical [2] divided into hierarchical methods, as complete-linkage, single-linkage and partitional methods, as ISODATA, K-MEANS.

Hierarchical clustering methods specifically aim for building a hierarchy of clusters, which are represented by dendrograms. However, these methods are limited to only small numbers of data values in practice. Classical methods compare all data values with each other and calculate a (dis-)similarity matrix. A growing number of data values leads for most hierarchical algorithms to non practical calculation and memory efforts.

 Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2016

doi:10.1016/j.procs.2016.05.430

Specifically partitional methods, as K-MEANS, need as starting point a "good guess" on the positions and number of the initial cluster centers. Clustering of the remaining data values is less resource consuming. However, if the initial decision was bad, also these methods become very computation intensive by calculating new cluster centers.

To cope with these problems we follow a solution approach by shifting the problem to a tractable solution space. We omit to look at each data value specifically but aggregate the data values into disjoint sets and cluster the data values based on their sets. For building these sets we use the index of a value space organizing multi-dimensional data structure.

Thus our general approach can be described basically by three steps:

1. Insert data values into a multi-dimensional data structure creating an index, which partitions value space into a set of boxes containing data values

2. Create density index of each box as ratio of number of data values to space volume of box

3. Cluster data values via their boxes by starting from box with highest density index and continuously adding additional neighbouring boxes.

The hierarchical Grid-Clustering [5] and BANG-clustering algorithms [6, 7] are based on this approach. Data values are inserted into a multi-dimensional data structure (the Grid file and the BANG (Balanced And Nested Grid) file) respectively, which organize the value space surrounding the data values into grid boxes by an index structure. Using this index structure now starting with grid boxes of high density index, i.e. many data values in relation to covered value space, these boxes are joined with other boxes based on their neighbour relation delivering a hierarchical dendrogram. This algorithm outperforms conventional algorithms dramatically in performance and memory usage.

In this paper we present a framework which allows to evaluate different data structures for the generation of a multi-dimensional grid structure grouping the data values into blocks. As first data structure to be evaluated we present the BANG file structure and show its feasibility as clustering index. The developed framework will be contributed as package to the WEKA software. WEKA (Waikato Environment for Knowledge Analysis) [8] is a well-known software tool providing various machine learning and data mining techniques and methods.

The paper is organized as follows: In the next section 2 we present the BANG file clustering method, which is the first implemented approach in our framework. The overall architecture of the envisioned framework and a justification of the BANG file implementation as hierarchical clustering method is delivered in section 3. The paper closes with the conclusion and a view on future work.

2 BANG File Clustering

Many Grid-Clustering algorithms show some characteristic problems in few specific situations (see [5]). These problems can be shortly summarized as

- for specific clusterings the memory requirements of the directory size increase overproportionally to the size of the data set, and

- the performance of the Grid-Clustering decreases with increasing dimensionality (number of pattern attributes) of the data set.

To overcome these drawbacks we proposed the BANG-Structure [6, 7], which is derived from the multi-dimensional BANG-File data structure [4]. This BANG-Structure shows not only better behavior for the listed problems, but also adapts better to clusterings in the value space.

Instead of comparing all data values with each other, the BANG-Structure organizes the value space containing the patterns. The data values represent patterns in a k-dimensional value space and are, in a first step, inserted into the BANG-Structure. These patterns are stored in the index preserving their topological distribution. In figure 1 a 2-dimensional value space with 10000 patterns and 70% of data clustered in 3 centers and in figure 2 the respective BANG index structure organizing the data values in a set of encasing rectangular *blocks* are shown. A block is rectangularly shaped and contains up to a defineable maximal number of p_{max} patterns. X $= (x_1, x_2, ..., x_n)$ is a set of n patterns and x_i is a pattern consisting of a tuple of k describing features $(p_{i_1}, p_{i_2}, ..., p_{i_k})$, where k is the number of dimensions. Using the block information of the BANG-Structure the algorithm identifies cluster center and clusters the patterns by an iterative neighbour algorithm accordingly to their surrounding blocks.

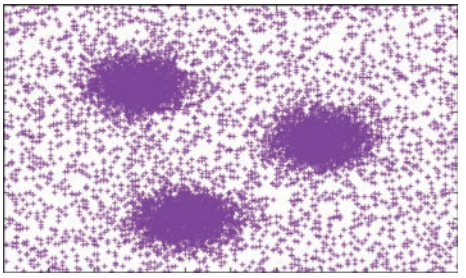

Figure 1: 2-dimensional value space example

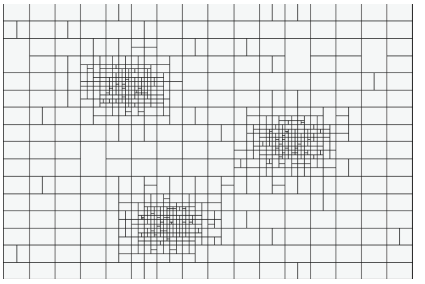

Figure 2: BANG partitioning

Figure 3: 2-dimensional pattern set - BANG clustered

3 Framework for Grid Clustering Methods

The BANG-Clustering algorithm was implemented in Java applying the MVC (Model-View-Controller) paradigm, clearly separating the user interaction from the business logic and the presentation. This allows to split the software package in concise but comprehensive components.

The general architecture of the Grid clustering framework consists of five software layers:

1. The *Data Management layer* administrates the data set for the clustering process. Hereby several formats are supported, as CVS, and access to database system is provided via an ODBC/JDBC interface.

2. The *Index Structure layer* allows to choose a specific data structure for the data space partitioning process. Until now only the BANG file is supported, but other data structure,as the R-tree, will be provided soon.

3. The *Clustering layer* uses the grid partitioning of the Index Structure layer and performs the agglomerative clustering by different neighbour search approaches.

4. The *Analysis layer* allows to inspect the computed clusters and to compare the results of clustering approaches using different index structures.

5. The *Presentation layer* is for the output of the clustering results in human friendly form, as dendrograms.

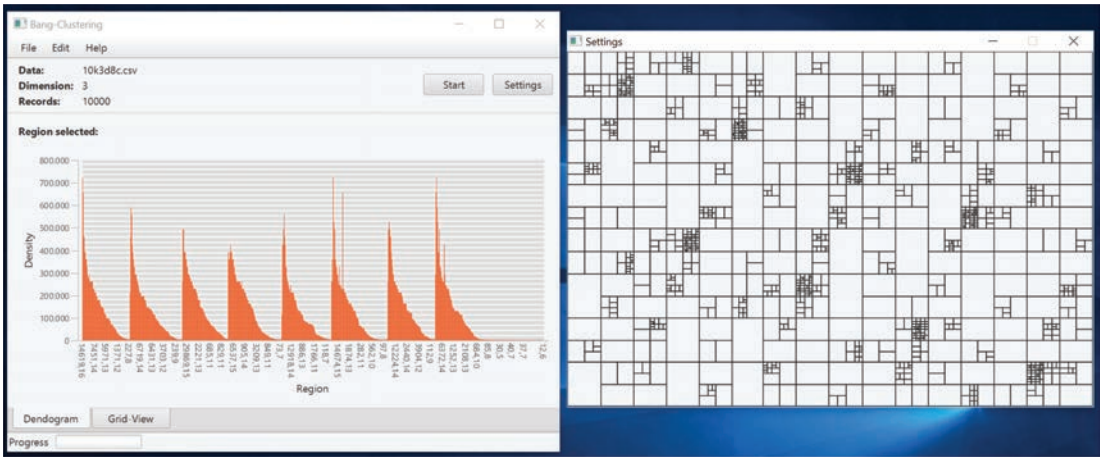

Figure 4: BANG system windows

In figure 4 the clustering approach is presented for a 3-dimensional data set consisting of 10000 patterns where 70% of data are grouped in 8 cluster centers. See Figure 5 for a 3D view. In the left widow of the system the final result in form of a dendrogram is shown. It can be easily seen that all 8 cluster centers are identified. In the right window the BANG cluster index is given, which shows the value space partitioning. This specific example is not easily to be analyzed with conventional approaches, due to its size and the specific location of the cluster centers. On the one hand due to the large number of data values classical hierarchical methods are computationally hardly tractable, on the other hand for partitioning methods first approximations on the position and number of cluster centers are difficult to "guess". Due to the position of the centers they are hiding to the eye of a data scientists in a projected view, as show in Figure 6.

The second version of the system is delivered as plugin package to be integrated into the WEKA system. WEKA is written in Java and is freely available under the GNU software license. WEKA is very common in the area of data mining. However it provides only limited tools for cluster analysis, as the k-means algorithm. These reasons made WEKA the perfect environment for our new method, which was realized as WEKA package. Thus, we chose Java as coding language and we followed during the development of our BANG clustering method a test-based approach to guarantee stability of our software.

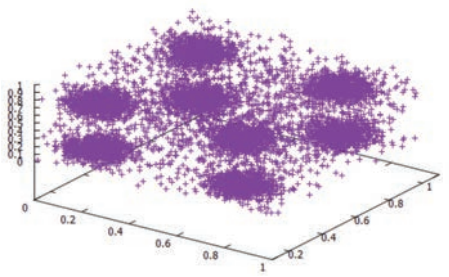

Figure 5: 3-dimensional example pattern set

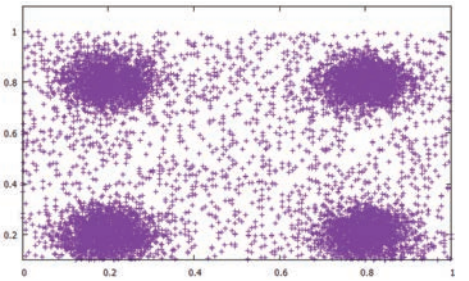

Figure 6: Data set projected to 2 dimensions

4 Conclusion and Future Work

In this paper we presented a framework for Grid-based cluster algorithms. This system allows to use and analyse different multi-dimensional index structures as basis for the clustering algorithm, which generate a partitioning of the value space by their index structure. In a final version the user will have the possibility to compare and choose between various index structures. Until now the BANG file structure is implemented, as presented in this paper. Several more multi-dimensional structures are in the focus of future research, as R-tree, kd-tree, etc. The already implemented system will be available in two versions: a stand-alone version with proprietary user interface and data manipulation and a plugin package, which integrates into the WEKA environmant.

References

[1] Pavel Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer, 2006.

[2] R. Dubes and A.K. Jain. *Clustering methodologies in exploratory data analysis*, volume 19, pages 113–228. Academia Press, 1980.

[3] Benjamin S Duran and Patrick L Odell. *Cluster analysis: a survey*, volume 100. Springer Science & Business Media, 2013.

[4] M.W. Freestone. The bang file: A new kind of grid file. In *Proc. Special Interest Group on Management of Data*, pages 260–269. ACM, May 1987.

[5] E. Schikuta. Grid clustering: An efficient hierarchical clustering method for very large data sets. In *Proc. 13th Int. Conf. on Pattern Recognition*, volume 2, pages 101–105. IEEE Computer Society, 1996.

[6] Erich Schikuta and Martin Erhart. The bang-clustering system: Grid-based data analysis. In X Liu, P. Cohen, and M. Berthold, editors, *Advances in Intelligent Data Analysis. reasoning about Data, Proc. Second International Symposium IDA-97*, volume 1280 of *LNCS*, London, UK, August 1997. Springer-Verlag.

[7] Erich Schikuta and Martin Erhart. Bang-clustering: A novel grid-clustering algorithm for huge data sets. In *Advances in Pattern Recognition*, pages 867–874. Springer, 1998.

[8] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[9] Rui Xu, Donald Wunsch, et al. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.